TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Informatik

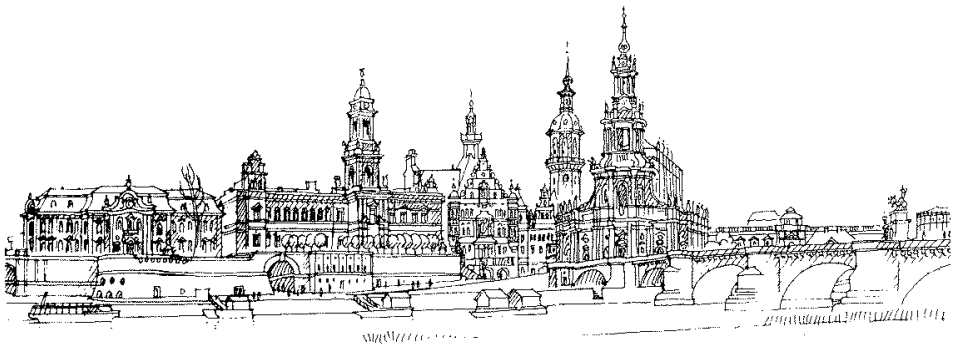
Technische Berichte
Technical Reports
ISSN 1430-211X

TUD-FI09-12-Dezember 2009

Eva-Maria Schwartz

Institut für Angewandte Informatik

Komparativer Ähnlichkeitsalgorithmus



Technische Universität Dresden
Fakultät Informatik
D−01062 Dresden
Germany
URL: http://www.inf.tu−dresden.de/

TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Informatik Institut für Angewandte Informatik, PDAI

KOMPARATIVER ÄHNLICHKEITSALGORITHMUS

Algorithmus zur komparativen Bewertung der Ähnlichkeiten von Objekten anhand von kollaborativen Priorisierungen

Dipl.-Inf. Eva-Maria Schwartz

INHALTSVERZEICHNIS

TABELLENVERZEICHNIS

Adresse
Fakultät Informatik
Institut für Angewandte Informatik (PDAI)
TU Dresden
01062 Dresden

Telefon: 0351 463 37897
Fax: 0351 463 38433
E-Mail: schwartz@pdai.de
http://inf.tu-dresden.de

1. MOTIVATION

Die „Nutzung der kollektiven Intelligenz“ ist der Hauptaspekt im Web 2.0. Das Wissen der Nutzer soll in die Anwendungen so integriert werden, dass andere Anwender davon profitieren können.

In den letzten Jahren wurden in den unterschiedlichsten Bereichen verschiedene Konzepte und Anwendungen entwickelt, welche dieses Prinzip umsetzen. Das wohl bekannteste Beispiel aus dem Web 2.0-Bereich ist das 2001 gegründete Wikipedia [Wik09]. Nach Angaben von Wikipedia wurden bisher über 900.000 Artikel veröffentlicht. Nutzer können Artikel anlegen und andere Artikel verändern bzw. ergänzen. Das bedeutet, dass die Inhalte nicht zentralisiert von Unternehmen erstellt und verbreitet werden, sondern von einer Vielzahl von Nutzern. In den letzten Jahren wurde das Prinzip auch im kommerziellen Geschäftsbereich eingesetzt, wie zum Beispiel das „Tchibo-ideas-Portal“, welches die Firma Tchibo GmbH am 12.05.2008 gründete [Tch09]. Das Konzept des Portals ist, Aufgaben und Lösungen für neuartige Produkte bzw. Alltagsprobleme zu finden. Damit öffnete Tchibo GmbH ihr Geschäftsmodell für die Außenwelt, um so an neue Ideen und Innovationen zu gelangen. Eine weitere Erfolgsgeschichte konnte Amazon schreiben. Das Online-Kaufhaus besteht seit Oktober 1998 und hat zum Ziel, das kundenorientierteste Unternehmen der Welt zu sein, bei dem Kunden alles finden, was sie online kaufen wollen. Amazon.de ist in den letzten elf Jahren zum bekanntesten Online-Shop geworden, wobei die Integration der Produktbewertungen durch Nutzer einen entscheidenden Marktvorteil verschafft. Grundsätzlich kann bei Amazon.de zwischen aktiven (der Nutzer gibt Meinungen und Bewertungen per Eingabe ab) und passiven (das Kauf- und Navigationsverhalten des Nutzers wird analysiert) Bewertungen unterschieden werden. Die entsprechenden Daten werden ausgewertet und als Empfehlungen in verschiedenster Art und Weise dargestellt. Diese Analysen und die daraus entstehenden Empfehlung werden durch ein Recommender-System[1] umgesetzt.

Das Prinzip der „Nutzung der kollektiven Intelligenz“ konnte bereits erfolgreich in Social-Software-Systemen und im Bereich E-Commerce eingesetzt werden. Aus diesem Grund werden jetzt Wege gesucht, um es in Software beim allgemeinen Arbeitseinsatz zu integrieren. Damit soll der Nutzer bei seiner täglichen Arbeit unterstützt werden. Dies ist besonders interessant und notwendig, wenn die Programme nicht wie im klassischen Vorgehen individuell auf die Bedürfnisse des Nutzers entwickelt (User-Centered-Design), sondern für einen möglichst breiten Nutzerkreis geschaffen und durch Konfigurationen angepasst werden.

1.1. AUSGANGSSITUATION

Die Notwendigkeit zur Nutzung von nicht-individuell entwickelter Software entsteht im Geschäfts- und Arbeitsfeld auf Grund der Entwicklung in diesem Bereich. Unternehmen müssen sich ständig ändernden Anforderungen im Geschäftsumfeld stellen. Mit dem immer stärker werdenden Wettbewerb ist es erforderlich, sich auf eigene Kernkompetenzen zu konzentrieren und zeitliche Kooperation bzw. Beziehungen mit anderen Organisationen

[1] Recommender-Systeme sind Werkzeuge zur Erstellung und Verbreitung von Empfehlungen. Der Sinn dieser Systeme ist nach [MR08], Informationen zu filtern, aufzubereiten und wertvolle Empfehlungen für den Benutzer zu geben.

einzugehen. Um diesen Beziehungen und Anforderungen gerecht zu werden, müssen Software bzw. Softwarebausteine flexibel und temporär bezogen werden.

Um den Nutzern dieser Software eine bestmögliche Unterstützung bei der Auswahl ihrer bedarfsgerechten Komponenten zu geben, sollen Ihnen, anhand von Entscheidungen bereits bestehender Kunden, Vorschläge für Objekte unterbreitet werden. Diese Objekte können je nach System zum Beispiel Konfigurationseigenschaften, Inhaltsmodule oder Layoutdarstellungen sein.

Es wird davon ausgegangen, dass ähnliche Nutzer auch ähnliche Objekte benötigen. Aus diesem Grund sollen die Nutzer miteinander verglichen werden. Das Problem liegt an dieser Stelle in der Beschreibung eines Nutzers. Dieser kann durch eine Vielzahl von Merkmalen gekennzeichnet werden, welche je nach Objekt eine unterschiedliche Wichtigkeit bei der Entscheidung haben. Aus diesem Grund müssen die einzelnen Merkmale unabhängig von einander betrachtet werden. Bei der Bewertung eines Objektes sollen dann entsprechende Wichtungen für das jeweilige Merkmal integriert werden.

Der Vergleich ist erst dadurch möglich, dass der Kontext und damit die Aufgabe des Nutzers bekannt sind. Nur mit diesen Informationen können gezielte Empfehlungen erstellt werden.

Im kommenden Abschnitt wird ein Verfahren vorgestellt, welches die priorisierte Bewertung einzelner Merkmale einbezieht. Ausgehend von diesem Verfahren wird ein Algorithmus vorgestellt, welcher Nutzer anhand ihrer Merkmale vergleicht und daraus folgend Empfehlungen für Objekte ausgibt. Der Algorithmus soll in ein Recommender-System integriert werden.

1.2. KOMPARATIVES BEWERTEN

Zur Lösung des Problems der polyoptimalen Ablaufsteuerung in der Fertigungssteuerung wurde in [Zfk76] ein Prioritätsalgorithmus vorgestellt, welcher die priorisierte Bewertung einzelner Merkmale einbezieht. Prioritätsalgorithmen sind heuristische Prinzipen zur Lösung dieses Problems. Das vorgestellte Verfahren zur komparativen Bewertung lässt sich in zwei Klassen einteilen:

1. Verfahren der Punktbewertung
2. Verfahren des diskreten, relativen Vergleiches

Ziel des Verfahrens ist es, Objekte von verschiedenen Beurteilenden bewerten zu lassen. Dabei werden die einzelnen vorher festgelegten Merkmale der Objekte separat bewertet.

In dieser Arbeit wird nur auf die Klasse zum Verfahren der Punktbewertung eingegangen.

VERFAHREN ZUR PUNKTBEWERTUNG

Beim Verfahren der Punkbewertung werden ausgewählten Merkmalen jedes Objektes, durch einen oder mehrere Beurteilende, elementare Punktbewertungen zugeordnet. Die Gesamtheit der elementaren Punktbewertungen der ausgewählten Merkmale durch die Gesamtheit der Beurteilenden wird über eine geeignet festegelegte Gewichtsfunktion zur resultierenden komparativen Gesamtbewertung zusammengefasst.

Eine Menge von Objekten H = {h} mit card(h) = m0 wird beschrieben durch eine Menge M von Merkmalen. Das Merkmal kann dabei komplex qualitativ oder quantitativ sein.

$$H \subseteq \prod_{k=1}^{n} P(M_k) = M$$

Dabei gilt card(H) = m_0 - Anzahl der Verfahren

und dim(M) = n - Anzahl der Merkmale.

Die gegebene Menge von Objekten soll durch l Beurteilende $b \in B$ (mit card(B) = l) beurteilt werden.

Tabelle 1: Bezeichnungen des komparativen Bewertens

Bezeichnung	Element	Anzahl der Elemente	Index eines Elements
Beurteilender	$b \in B$	card(B) = l	i
Objekt	$h \in H$	card(H) = m0	j
Merkmal	$M_k \in M$	dim(M) = n	k

Der Algorithmus besteht aus folgenden 4 Schritten:

Schritt 1 Bestimmung einer Teilmenge $HT_1 \subseteq H$ von Objekten, welche eine gegebene Aufgabenstellung oder Norm erfüllen.

Schritt 2 Der i-te Beurteilende ordnet für das j-te Objekt ($\forall h \in HT_1$) dem k-ten Merkmal einen Wert („Bewertung“) $w_{i,j,k}$ zu. Die Bewertungen sind dabei festgelegte Wertungen mit vorgegebener Kardinalität. Für jedes Merkmal werden Gewichtsfaktoren festgelegt. Durch diese Faktoren g_k wird jedem j-tem Objekt durch einen i-ten Beurteilenden eine Bewertung wi,j zugeordnet

$$wi, j = \sum_{k=1}^{n} g_k * w_{i,j,k}$$

Durch diese Bewertung wird von jedem i-ten Beurteilenden die Menge der Objekte halbgeordnet. Dadurch ist dem j-ten Objekt eine Beurteiler-Platzzahl $p_{i,j}$ eindeutig zugeordnet mit

$$1 \le p_{i,j} \le card(HT_1) = m$$

$$\sum_{j=1}^{m} p_{i,j} = 1 + 2 + ... + m \text{ für } \forall i$$

Schritt 3 Für jedes der m Objekte werden die Beurteiler-Platzzahlen $p_{i,j}$ zu einer Platzzahl p_j unter Berücksichtigung von festgelegten Gewichtsfaktoren f_i der Beurteiler zusammengefasst:

$$p_j = \sum_{j=1}^{m} f_i * p_{i,j}$$

Schritt 4 Als beste Objekte werden diejenigen Objekte angesehen, die Element einer Teilmenge $HT_2 \subseteq HT_1$ sind, deren Platzzahl den kleinsten Wert hat:

$$HT_2 = \{h_k \in HT_1 \mid p_j = \inf\}$$

BEISPIEL

Zu einer Konferenz wird eine Vielzahl von fachlichen Beiträgen eingereicht. Da nicht alle Beiträge angenommen werden, soll jeder Beitrag von drei unterschiedlichen Beurteilenden

bewertet werden. Dabei sollen folgende Merkmale der eingereichten Beiträge beurteilt werden:

- Originalität
- Qualität
- Relevanz zum Thema der Konferenz
- Präsentation
- Empfehlung zur Annahme

Jeder Beurteilende vergibt für diese Merkmale Punkte nach dem Schulnotensystem. Dabei gilt 1 als die beste und 6 als die schlechteste Wertung. Eine mögliche Verteilung von zwei Beiträgen ist in Tabelle aufgezeichnet.

Tabelle 2: Beispieldaten für das komparative Bewerten

	Merkmale/ Beurteilende	Originalität	Qualität	Relevanz	Präsentation	Empfehlung
Beitrag 1	Bewerter 1	1	2	2	3	2
	Bewerter 2	2	1	1	2	1
Beitrag 2	Bewerter 1	1	2	1	3	3
	Bewerter 2	1	2	1	2	2

Ohne eine Priorisierung von Merkmalen entsteht folgendes Ergebnis:

Tabelle 3: Auswertung der Beispieldaten ohne Priorisierung

	Punktzahl Beitrag 1	Punktzahl Beitrag 2	Platzangabe Beitrag 1	Platzangabe Beitrag 2
Bewerter 1	10	10	1	1
Bewerter 2	7	8	1	2

Damit wird im direkten Vergleich Beitrag 1 angenommen.

Für die Veranstalter der Konferenz ist jedoch besonders die Originalität und Relevanz der Beiträge wichtig. Aus diesem Grund werden diese Merkmale mit dem Faktor 2 bewertet, während die anderen Merkmale mit dem Faktor 1 bewertet werden.

Tabelle 4: Auswertung der Beispieldaten mit priorisierten Merkmalen

	Punktzahl Beitrag 1	Punktzahl Beitrag 2	Platzangabe Beitrag 1	Platzangabe Beitrag 2
Bewerter 1	13	10	2	1
Bewerter 2	12	10	2	1

Nach der Priorisierung von Merkmalen wird Beitrag 2 im direkten Vergleich angenommen.
Im Weiteren können Bewerter ihr eigenes Wissen zu dem beschriebenen Gebiet einschätzen. Dabei gibt Bewerter 1 an, dass er ein fundiertes Wissen hat, während

Bewerter 2 nur grundlegende Kenntnisse besitzt. Aus diesem Grund können im nachfolgenden Schritt die Ergebnisse noch einmal priorisiert werden. Zum Beispiel könnten die Ergebnisse von Bewerter 1 doppelt so hoch eingehen wie die von Bewerter 2. In diesem Beispiel würde das jedoch nichts an der Platzangabe ändern.

1.3. GRUNDSÄTZLICHES VORGEHEN

Auf Grund des beschriebenen Verfahrens des komparativen Bewertens ist die Idee entstanden, die Ansätze des Verfahrens bei der Empfehlung für Objekte für Nutzer anhand von Nutzermerkmalen zu integrieren. Das Ziel dabei ist, Nutzern Objekte zu empfehlen, welche andere Nutzer mit ähnlichen Merkmalen bereits benutzen. Durch diese Empfehlungen kann die kollektive Intelligenz der Systemnutzer angewendet werden. Grundsätzlich sollen die Merkmale der Nutzer miteinander verglichen werden, wobei eine geeignete Priorisierung erfolgt, und die Objekte, welche die Nutzer mit der höchsten Ähnlichkeit verwenden, angezeigt werden. Dabei wird davon ausgegangen, dass die einzelnen Merkmale des Nutzers nicht immer gleich priorisiert werden, sondern in Abhängigkeit vom Objekt gewichtet werden müssen. Diese Wichtungsfaktoren müssen aus der Masse der bestehenden Daten (Nutzer und deren angewendete Objekte) analysiert und berechnet werden. Die periodisch zu berechneten Wichtungen können als kollaborative Priorisierungsfaktoren angesehen werden.

2. BESCHREIBUNG DER ALLGEMEINEN AUFGABENSTELLUNG

2.1. AUSGANGSSITUATION

Gegeben sei ein Objekt O mit den Eigenschaften E_1 bis E_S. Jede dieser Eigenschaften kann unterschiedliche Ausprägungen besitzen. Die Eigenschaften des Objekts sind voneinander unabhängig und die einzelnen Ausprägungen der Eigenschaften können unabhängig ausgewählt werden. Jede beliebige Auswahl (e_1, ..., e_S) der Eigenschaften E_1 bis E_S wird als Objektkonfiguration bezeichnet.

Jeder Nutzer des Objekts O bevorzugt eine für ihn individuelle Konfiguration. Die für den Nutzer geeignete Einstellung hängt von seinen individuellen Merkmalen ab. Der Nutzer wird durch eine endliche Anzahl von Merkmalen M_1 bis M_N beschrieben. Jedes dieser Merkmale kann unterschiedlich erfassbare Ausprägungen besitzen. Damit kann der Nutzer durch den Merkmalsvektor (m_1, ..., m_N) beschrieben werden.

Vorausgesetzt wird, dass Nutzer mit einem identischen Merkmalsvektor die gleiche Objektkonfiguration bevorzugen.

BEISPIEL

Die Nutzer werden mit folgenden Merkmalen und Attributen beschrieben:

1. Alter
   - 1 für 18 bis 30 Jahre
   - 2 für 31 bis 50 Jahre
   - 3 für 51 bis 70 Jahre
   - 4 für 71 bis 100 Jahre

2. Geschlecht
   - 1 für männlich
   - 2 für weiblich
3. Aufgabe
   - 1 für Geschäftsführer
   - 2 für Kaufmännischer Leiter
   - 3 für Finanzbuchhaltung
   - 4 für Lagerarbeiter
4. Zugangsgerät
   - 1 für Desktop-Rechner
   - 2 für Laptop
   - 3 für Mobiltelefon
   - 4 für Smartphone
5. Interaktionsmedium
   - 1 für Maus
   - 2 für Tastatur
   - 3 für Finger
   - 4 für Stick

Die Anzahl der Merkmale ist 5 und damit ist N=5. Der Vektor kann folgerdermaßen dargestellt werden:

$$\overrightarrow{M} = \begin{pmatrix} M_1 - Alter \\ M_2 - Geschlecht \\ M_3 - Aufgabe \\ M_4 - Zugangsgerät \\ M_5 - Interaktionsmedium \end{pmatrix}$$

Ein 52jähriger Mann, welcher im Lager die Pakete den einzelnen Lastern zuweist und dazu ein Smartphone verwendet, welches er ausschließlich per Touchscreen und seinen Fingern verwendet, kann so beschrieben werden:

$$\overrightarrow{M} = \begin{pmatrix} 3 \\ 2 \\ 4 \\ 4 \\ 5 \end{pmatrix}$$

2.2. BESONDERE ANFORDERUNGEN

Anforderung 1: Eine Ähnlichkeit zwischen zwei Nutzern und deren verwendeten Objekten kann nur analysiert werden, wenn die Merkmale bei beiden sinnvoll gewählt wurden.

Anforderung 2: Der Begriff „ähnlich“ muss näher beschrieben und definiert werden.

Anforderung 3: Die Merkmale eines Nutzers können von verschiedener Qualität sein. Es wird mindestens unterschieden zwischen:

- Nominalen Merkmalen: Merkmale, die nur Qualitäten beschreiben aber keine Ordnungskriterien besitzen
- Ordinale Merkmale: Merkmale, die eine natürliches oder abgebildetes Ordnungskriterium besitzen.

Anforderung 4: Die Merkmale des Nutzers haben unterschiedliche Wertungen (Prioritäten), welche bei unterschiedlichen Objekteigenschaften gänzlich verschieden sein können.

2.3. ABLEITBARE THESEN

Ausgehend von der beschriebenen Aufgabenstellung wird von folgenden Thesen ausgegangen:

These 1: Nützliche Informationen können aus der bekannten Nutzergruppe gewonnen werden.

These 2: Ein neuer Nutzer mit dem Merkmalsvektor $M_0=(m_{10}, ..., m_{N0})$ benutzt die gleiche Objektkonfiguration wie ein bereits bestehender Nutzer mit dem gleichen Merkmalsvektor.

These 3: Nutzer mit ähnlichen Merkmalen bevorzugen ähnliche Konfigurationen.

These 4: Bei ordinalen Werten kann man den Begriff „ähnlich“ mit einer geeigneten Metrik (Abstandsfunktion) d beschreiben.

These 5: Bei nominalen Merkmalen kann der Begriff „ähnlich“ durch „gleich“ oder „ungleich“ ersetzt werden.

$$d(a,b)=\begin{cases}0\ wenn\ a\equiv b\\1\ wenn\ a\neq b\end{cases}$$

These 6: Die Prioritäten der Nutzermerkmale in Zusammenhang mit den Objekten können anhand der bestehenden Daten ermittelt werden.

3. MATHEMATISCHE FORMULIERUNG

3.1. BESCHREIBUNG DER OBJEKTEIGENSCHAFTEN UND DEREN ATTRIBUTE

Der Vektor $\vec{E}=\begin{pmatrix}e_1\\e_2\\\vdots\\e_{s-1}\\e_s\end{pmatrix}$ beschreibt die Objekteigenschaften.

Jede Systemeigenschaft E_k besitzt eine endliche Anzahl von Attributen n_k, die durch den Vektor $a_k = (a_{k1}, a_{k2}, \ldots, a_{kn_k})$ beschrieben werden. Die Attribute können nominal oder ordinal sein.

3.2. BESCHREIBUNG DER NUTZER UND DEREN ATTRIBUTE

Ein Nutzer wird durch einen n-dimensionalen Vektor von Merkmalen M_1 bis M_n charakterisiert. Er wird folgendermaßen bezeichnet: $\overrightarrow{M} = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_{n-1} \\ m_n \end{pmatrix}$ Die einzelnen Merkmale M_k werden dabei endlich vielen Attributen (Ausprägungen, Werte) zugeordnet. Die Anzahl der Attribute kann unterschiedlich sein und wird als l_k für das Merkmal M_k bezeichnet.

3.3. WICHTUNG VON MERKMALEN

Bei der Erarbeitung von Empfehlungen werden im Allgemeinen alle Merkmale des Nutzers eine Rolle spielen. Jedoch können die einzelnen Merkmale mit unterschiedlichem Gewicht in die Entscheidungsfindung einfließen.

Für die n Merkmale des Nutzers wird deshalb ein n-dimensionaler Wichtungsvektor eingeführt. $\overrightarrow{W} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{n-1} \\ w_n \end{pmatrix}$ Diese Wichtungsfaktoren sind allerdings nicht a-priori bekannt. Sie müssen im Laufe eines Verfahrens permanent aus den bekannten Werten mitberechnet und gegebenenfalls angepasst werden.

Die Wichtungsfaktoren werden im Allgemeinen bei den einzelnen Objekteigenschaften E_1 bis E_n unterschiedlich sein und sind damit für jede Systemeigenschaft gesondert zu bestimmen. Für alle Wichtungsfaktoren w_i gelte $0 \leq w_i \leq 1$. Ist eine a-priori Schätzung für die Wichtungsfaktoren möglich, so kann diese als Anfangswichtung benutzt werden. Ist keine a-priori Schätzung möglich, werden anfänglich alle Wichtungsfaktoren w_i auf 1 gesetzt.

3.4. ANFANGSINFORMATIONEN

Um Empfehlungen bestimmen zu können, wird eine ausreichende Anzahl von Anfangsinformationen benötigt. Es wird vorausgesetzt, das N Nutzer vorhanden sind, deren

Merkmale und ihre Entscheidung für eine Objektkonfiguration bekannt sind. Diese Daten werden in folgender Form beschrieben:

NUTZERMATRIX N

In der (n,N)-dimensionalen Nutzermatrix $N = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1N-1} & m_{1N} \\ m_{21} & m_{22} & \cdots & m_{2N-1} & m_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{n-11} & m_{n-12} & \cdots & m_{n-1N-1} & m_{n-1N} \\ m_{n1} & m_{n2} & \cdots & m_{nN-1} & m_{nN} \end{pmatrix}$

werden die Merkmalswerte der N Nutzer zusammengefasst. Dabei beschreibt das Element m_{ij}, dass in der i-ten Zeile und j-ten Spalte steht, den i-ten Merkmalswert des j-ten Nutzers.

ERGEBNISMATRIX E

In der (S,N)-dimensionalen Ergebnismatrix $E = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1N-1} & e_{1N} \\ e_{21} & e_{22} & \cdots & e_{2N-1} & e_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_{S-11} & e_{S-12} & \cdots & e_{S-1N-1} & e_{n-1N} \\ e_{S1} & e_{S2} & \cdots & e_{SN-1} & e_{SN} \end{pmatrix}$ werden

die Entscheidungen der n Nutzer für ihre Objekteigenschaften zusammengefasst. Dabei beschreibt das Element in e_{ij}, dass in der i-ten Zeile und j-ten Spalte steht, den i-ten Attributswert des j-ten Nutzers.

WICHTUNGSMATRIX W

Die Wichtungsfaktoren werden in der (n, S) dimensionalen Wichtungsmatrix

$$W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1S-1} & w_{1S} \\ w_{21} & w_{22} & \cdots & w_{2S-1} & w_{2S} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_{n-11} & w_{n-12} & \cdots & w_{n-1S-1} & w_{n-1S} \\ w_{n1} & w_{n2} & \cdots & w_{nS-1} & w_{nS} \end{pmatrix}$$

zusammengefasst. Das Element w_{ij}, welches in der i-ten Zeile und j-ten Spalte steht, beschreibt dabei den Wichtungsfaktor des i-ten Attributwertes des j-ten Nutzers.

ÄHNLICHKTEITSFUNKTIONEN

Um einen neuem (aktuellen) Nutzer eine Empfehlung auf der Basis der vorhandenen Informationen (Nutzermatrix und Ergebnismatrix) zu geben, ist es wünschenswert in der Nutzermatrix „ähnliche Nutzer“ zu finden und deren Entscheidungen als Grundlage für die

Empfehlung bzw. Empfehlungen zu verwenden. Dazu muss der Begriff „Ähnlichkeit“ für die einzelnen Merkmale mathematisch beschrieben werden.

Grundsätzlich muss eine Ähnlichkeitsfunktion f(x,y) die folgenden Eigenschaften haben :

f(x,x) = 0 für alle x

f(x,y) > 0 für alle $x \neq y$

Wünschenswert ist weiterhin, $f(x, y_2) > f(x, y_1)$, wenn der „Unterschied“ zwischen x und y_2 „größer“ ist als zwischen x und y_1.

Ist das Merkmal nominal, dann bietet sich als Ähnlichkeitsfunktion die Funktion

$$f(x,y) = \begin{cases} 0 & \textit{für } x = y \\ 1 & \textit{für } x \neq y \end{cases}$$

an. Ist das Merkmal ordinal, dann bieten sich als Ähnlichkeitsfunktion die Funktionen

$$f(x,y) = | y - x |$$

bzw.

$$f(x,y) = (x-y)^2$$

an. In beiden Fällen können die Funktionen natürlich auch noch mit geeigneten Gewichtsfaktoren multipliziert werden.

Im Weiteren wird vorausgesetzt, dass die Funktionen $f_1, \ldots, f_n$ für die n Merkmale des Nutzers geeignete Ähnlichkeitsfunktionen sind.

ÄHNLICHKEITSMATRIX

Der Merkmalsvektor des neuen (aktuellen) Nutzers sei

$$M_0 = \begin{pmatrix} m_{10} \\ m_{20} \\ \vdots \\ m_{n0} \end{pmatrix}.$$

In der (n,N)-dimensionalen Ähnlichkeitsmatrix

$$F = \begin{pmatrix} f_1(m_{10}, m_{11}) & f_1(m_{10}, m_{12}) & \cdots & f_1(m_{10}, m_{1N}) \\ f_1(m_{20}, m_{21}) & f_1(m_{20}, m_{22}) & \cdots & f_1(m_{20}, m_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(m_{n0}, m_{n1}) & f_1(m_{n0}, m_{n2}) & \cdots & f_1(m_{n0}, m_{nm}) \end{pmatrix}$$

werden die Ähnlichkeitswerte des aktuellen Nutzers und der N bekannten Nutzer zusammengefasst. Dabei beschreibt das Element, das in der i-ten Zeile und j-ten Spalte steht, den i-ten Ähnlichkeitswert des aktuellen Nutzers mit dem j-ten bekannten Nutzer.

4. BESCHREIBUNG DES ALGORITHMUS

Der Algorithmus gliedert sich in zwei Teile:

1. Bestimmung eines „ähnlichen“ Nutzers und der Entscheidung, worauf aufbauend Empfehlung bzw. Empfehlungen getroffen werden.
2. Algorithmus zur sukzessiven Verbesserung der Wichtungsfaktoren während der gesamten Laufzeit des Systems.

Diese jeweiligen Algorithmen werden in den kommenden Abschnitten erläutert.

4.1. AUFFINDEN EINES „ÄHNLICHEN“ NUTZERS

VORAUSSETZUNG

Für eine formale Beschreibung müssen folgende Anforderungen erfüllt sein.

Anforderung 1: Für alle Merkmale und Objekteigenschaften muss eine Zuordnung zu den natürlich Zahlen gewählt werden. Dazu muss eine geeignete Klasseneinteilung der Merkmals- und Objektausprägungen festgelegt werden. Dies ist nicht nötig, wenn bereits eine natürlicherweise vorhanden ist.

Anforderung 2: Sind die Ausprägungen nominal, werden sie durchnummeriert. Bei ordinalen Ausprägungen wird die Ordnung mit übernommen.

Anforderung 3: Alle Merkmalseigenschaften der schon bekannten Nutzer sind erfasst.

Anforderung 4: Die Objekteigenschaften des benutzten Objektes eines jeweiligen Nutzers sind erfasst.

FORMALER ALGORITHMUS

Tabelle 5: Bezeichnungen des komparativen Ähnlichkeitsalgorithmus

Bezeichnung	Element	Anzahl der Elemente	Bezeichnung der Attribute
Bekannten Nutzer	N	N	
Merkmales des Nutzers	M	n	l
Eigenschaften des Objektes	E	S	a

Gegeben sind folgende Werte:

Die Nutzermatrix N(n,N), die dazugehörige Ergebnismatrix E(S,N) und die aktuelle Wichtungsmatrix W(n, S) seien bekannt.

Als Abstandsfunktion wird gewählt:

$$d_i(x,y) = \begin{cases} \begin{cases} 0 & wenn \quad x \neq y \\ & \text{wenn} \quad \mathrm{M_i} \quad \text{nominal} \\ 1 & wenn \quad x = y \end{cases} \\ |x-y| \qquad \text{wenn} \quad \mathrm{M_i} \quad \text{nominal} \end{cases}$$

Sei $M_0 = \begin{pmatrix} m_{10} \\ m_{20} \\ \vdots \\ m_{n0} \end{pmatrix}$ der Merkmalsvektor eines neuen Nutzers.

Berechnung der Ähnlichkeitsmatrix

Aus den gegebenen Werten kann nun die Ähnlichkeitsmatrix F mit

$$f(m_{io}, m_{ik}) = d_i(m_{io}, m_{ik}) \quad für \quad i = 1,\ldots,n \; und \; j = 1,\ldots,N$$

berechnet werden. Dazu müssen folgende Schritte ausgeführt werden:

Schritt 1 Setze i=1 (i-te Objekteigenschaft)

Schritt 2 Sei W_i die i-te Spalte der Wichtungsmatrix W. Berechne den Vektor $B = (b_1, b_2, \ldots, b_n)$ aus $B = W_i^T \bullet F$

Schritt 3 Sortiere die Komponenten des Vektors B aufsteigend. $B_{sort} = (b_{i_1}, b_{i_2}, \ldots, b_{i_n})$ Die Indexfolge $(i_1, i_2, \ldots, i_n)$ gibt nun die Reihenfolge der „ähnlichsten" Nutzer bezüglich des neuen Nutzers und der Objekteigenschaft i an.

Schritt 4 Fixiere die besten Objekteigenschaften aus der Ergebnismatrix E

Schritt 5 Wenn i > S dann i=i+1 und beginne mit Schritt 1 wieder, sonst Schritt 6

Schritt 6 Ausgabe der besten Objektkonfigurationen als Empfehlung.

Schritt 7 Aufnahme des neuen Nutzers in die Nutzermatrix und Erhöhung von N auf N+1

4.2. PERIODISCHE BERECHNUNG DER WICHTUNGSFAKTOREN

VORAUSSETZUNG

Anforderung 1: Allen Merkmalseigenschaften und Objekteigenschaften werden Werte von 1,2, ..., n zugeordnet.

FORMALE BESCHREIBUNG

Für alle 1 bis S Objekteigenschaften wird folgende Berechnung durchgeführt:

Sei E_i die gewählte Objekteigenschaft und M_j das j-te Nutzermerkmal

Schritt 1 Ordne jeder Merkmalsausprägung von m_j die zugehörige Objekteigenschaft zu.

Schritt 2 Bestimme den Mittelwert und die Varianz für jede Merkmalsausprägung.

Schritt 3 Addiere alle Varianzen auf. Die Summe sei Sum.

Schritt 4 Setze den Wichtungsfaktor w_{ij}, welches den Wichtungsfaktor des j-ten Merkmals in der i-ten Objekteigenschaft beschreibt, auf

$$w_{ij} = \frac{1}{1 + Sum}$$

Alle Wichtungsfaktoren liegen zwischen 0 und 1. Merkmale, die eine geringere Streuung bezüglich der Objekteigenschaft haben, sind größer als die mit einer großen Streuung.

5. ZUSAMMENFASSUNG

In dieser Arbeit ist ein Algorithmus zur Empfehlung von Objekten anhand Ähnlichkeiten zwischen Nutzern beschrieben, dabei werden die verschiedenen Merkmale der Nutzer verglichen. Die Priorisierung der Merkmale erfolgt mit kollaborativen Wichtungsfaktoren. Diese werden auf Grundlage der bereits bestehenden Daten von Nutzern berechnet.

Mit Hilfe dieses Algorithmus, welcher Grundlage eines Recommender-Systems werden soll, können Objekte für Nutzer im Arbeitsalltag empfohlen werden. Die Besonderheit im Bereich dieser Business-Systeme liegt im vorhandenen Kontext. Der Nutzer eines solchen Systems hat eine Aufgabe, welche er zu erfüllen hat und anhand dieser bekannten Aufgabe können notwendige Objekte angeboten werden. Die kollaborativen Wichtungsfaktoren beziehen sich jeweils auf ein entsprechendes Objekt. Es wird davon ausgegangen, dass Nutzermerkmale je nach Objekt unterschiedlich hoch priorisiert werden müssen.

Das Ziel des Algorithmus und des darauf aufbauenden Recommener-Systems ist es, die Probleme, welche durch „non-user-centered"-Design entstandenen Produkte zu lösen. Dabei sind die Nutzerdaten, welche auf Grund der neuen Nutzungsmodelle vorhanden sind, die neue Wissensgrundlage. In diesen Daten sollen Gesetzmäßigkeiten zur Anwendung von Objekten im Zusammenhang mit speziellen Nutzermerkmalen erkannt werden.

In den folgenden Arbeiten muss der beschriebene Algorithmus getestet werden. Dabei muss im Speziellen festgestellt werden, ab wann die berechneten Gewichtsfaktoren zuverlässig in ihrer Aussage sind. Dies impliziert eine Anzahl von vorhandenen Nutzerdaten

für das jeweilige Objekt. Die genaue Anzahl kann nur durch direkte Tests an speziellen Objekten bestimmt werden. Dabei wird die Auswahl der Merkmale von Nutzer und Objekten eine wichtige Rolle spielen.

6. LITERATUR

[Ama09] http://www.amazon.de, Zugriff: 15.05.2009

[MR08] Mürzl, G.; Riemenschneider H.: Recommender System der anderen Art: Kaufempfehlungen für Supermarktartikel. http://www.muerzl.net/data/preis_empfehlung.pdf, Abruf am 03.05.2008

[ST08] Schachner, Werner ; Tochtermann, Klaus: corporate web 2.0 band II. Shaker Verlag, 2008

[Tch09] https://www.tchibo-ideas.de, Zugriff: 15.05.2009

[Wik09] http://de.wikipedia.org, Zugriff: 16.05.2009

[ZfK76] Baldeweg, Frank; Jungcluassen, Hardwin; Stahn, Heinz: Grundlagen der Kybernetik II; Zentraliinstitut für Kernforschung Rossendorf bei Dresden, März 1976